CLASSIFICATION EXERCISE SET

Exercise 1:

Given the training data below, use a naïve Bayes classifier with *m*-estimate constant m=2 to the test patterns corresponding to the following people: a person who is coughing and has fever, a person whose nose is running and who suffers fever, and a person with a running nose and reddened skin.

Example	x_1 (running nose)	x_2 (coughing)	x_3 (reddened skin)	x_4 (fever)	Class
1	Yes	Yes	Yes	No	Positive (ill)
2	Yes	Yes	No	No	Positive (ill)
3	No	No	Yes	Yes	Positive (ill)
4	Yes	No	No	No	Negative (healthy)
5	No	No	No	No	Negative (healthy)
6	No	Yes	Yes	No	Negative (healthy)

Exercise 2:

Given the training data below, use a naïve Bayes classifier with *m*-estimate constant m=4 to the test patterns corresponding to the following people: a person with high fever, and a person who suffers from vomiting and shivering.

Example	x_1 (fever)	x_2 (vomiting)	x_3 (diarrhea)	x_4 (shivering)	Class
1	no	no	no	no	healthy
2	average	no	no	no	influenza
3	high	no	no	yes	influenza
4	high	yes	yes	no	salmonella poisoning
5	average	no	yes	no	salmonella poisoning
6	no	yes	yes	no	bowel inflammation
7	average	yes	yes	no	bowel inflammation

Exercise 3:

Given the training data below, use a naïve Bayes classifier with *m*-estimate constant m=5 to the test pattern corresponding to a student with medium income, and fair credit rating.

Example	x_1 (income)	x_2 (student)	x_3 (credit rating)	Class (buys a computer)
1	High	No	Fair	No
2	High	No	Excellent	No
3	High	No	Fair	Yes

4	Medium	No	Fair	Yes
5	Low	Yes	Fair	Yes
6	Low	Yes	Bad	No
7	Low	Yes	Excellent	Yes
8	Medium	No	Bad	No
9	Low	Yes	Fair	Yes
10	Medium	Yes	Bad	Yes
11	Medium	Yes	Excellent	Yes
12	Medium	No	Excellent	Yes
13	High	Yes	Fair	Yes
14	Medium	No	Excellent	No

Exercise 4:

Let us consider a medical diagnosis task. We know that about 10.6% of the overall population have diabetes. There is a laboratory test which returns a correct positive result in 99% of the cases in which the disease is present, and a correct negative result in 94% of the cases where the disease is not present.

a) Suppose we observe a patient such that the test returns positive. Compute the probability that he/she suffers from diabetes.

b) A second test (which is assumed to be independent of the first one) is carried out, and it returns a negative result. Recompute the probability that he/she suffers from diabetes.

Exercise 5:

Compute the following performance measures from the confusion matrix given below:

a) Accuracy.
b) True positives, true negatives, false positives, false negatives.
c) Precision, fallout, recall, F-measure.

	Actual class		
Predicted class		Positive	Negative
	Positive	182	31
	Negative	12	245

POSSIBLE SOLUTIONS:

Exercise 1:

In this exercise C=2, V={*Ill*, *Healthy*}, and D=4. Consequently, for every $y \in V$ we have:

$$P(y \mid \mathbf{x}) = \frac{P(y)P(\mathbf{x} \mid y)}{P(Ill)P(\mathbf{x} \mid Ill) + P(Healthy)P(\mathbf{x} \mid Healthy)}$$

$$P(\mathbf{x} \mid y) = \prod_{d=1}^{4} P(x_d \mid y)$$

From the class frequencies in the training set we obtain the class prior probabilities:

$$P(Ill) = \frac{3}{6} = \frac{1}{2}$$

$$P(Healthy) = \frac{3}{6} = \frac{1}{2}$$

On the other hand, since all the input components have only 2 possible values, we can set the a priori probabilities for the m-estimate to p=0.5, which means that:

$$P(x_d \mid y) = \frac{n'+1}{n+2}$$

Next we compute the posterior probabilities:

$P(x_1 = Yes \mid Ill) = \frac{2+1}{3+2} = \frac{3}{5}$ (we consider examples 1, 2 and 3)

$P(x_1 = Yes \mid Healthy) = \frac{1+1}{3+2} = \frac{2}{5}$ (we consider examples 4, 5 and 6)

$P(x_2 = Yes \mid Ill) = \frac{1+1}{3+2} = \frac{2}{5}$ (we consider examples 1, 2 and 3)

$P(x_2 = Yes \mid Healthy) = \frac{1+1}{3+2} = \frac{2}{5}$ (we consider examples 4, 5 and 6)

$P(x_3 = Yes \mid Ill) = \frac{2+1}{3+2} = \frac{3}{5}$ (we consider examples 1, 2 and 3)

$P(x_3 = Yes \mid Healthy) = \frac{1+1}{3+2} = \frac{2}{5}$ (we consider examples 4, 5 and 6)

$P(x_4 = Yes \mid Ill) = \frac{1+1}{3+2} = \frac{2}{5}$ (we consider examples 1, 2 and 3)

$P(x_4 = Yes \mid Healthy) = \frac{0+1}{3+2} = \frac{1}{5}$ (we consider examples 4, 5 and 6)

Finally, we can compute the probabilities that each person is ill:

a) A person who is coughing and has fever: **x**=(No,Yes,No,Yes).

$$P(\mathbf{x} \mid Ill) = P(x_1 = No \mid Ill)P(x_2 = Yes \mid Ill)P(x_3 = No \mid Ill)P(x_4 = Yes \mid Ill) =$$

$$\frac{2}{5}\cdot\frac{2}{5}\cdot\frac{2}{5}\cdot\frac{2}{5}=\frac{16}{625}$$

$$P(\mathbf{x}\mid Healthy)=$$

$$P(x_1=No\mid Healthy)P(x_2=Yes\mid Healthy)P(x_3=No\mid Healthy)P(x_4=Yes\mid Healthy)=$$

$$\frac{3}{5}\cdot\frac{2}{5}\cdot\frac{3}{5}\cdot\frac{1}{5}=\frac{18}{625}$$

$$P(Ill\mid\mathbf{x})=\frac{P(Ill)P(\mathbf{x}\mid Ill)}{P(Ill)P(\mathbf{x}\mid Ill)+P(Healthy)P(\mathbf{x}\mid Healthy)}=$$

$$\frac{\frac{1}{2}\cdot\frac{16}{625}}{\frac{1}{2}\cdot\frac{16}{625}+\frac{1}{2}\cdot\frac{18}{625}}=\frac{16}{16+18}=\frac{8}{17}\approx 0.4706$$

$$P(Healthy\mid\mathbf{x})=1-P(Ill\mid\mathbf{x})=\frac{9}{17}\approx 0.5294$$

Consequently, this person is predicted to be healthy.

b) A person whose nose is running and who suffers fever: **x**=(Yes,No,No,Yes).

$$P(\mathbf{x}\mid Ill)=P(x_1=Yes\mid Ill)P(x_2=No\mid Ill)P(x_3=No\mid Ill)P(x_4=Yes\mid Ill)=$$

$$\frac{3}{5}\cdot\frac{3}{5}\cdot\frac{2}{5}\cdot\frac{2}{5}=\frac{36}{625}$$

$$P(\mathbf{x}\mid Healthy)=$$

$$P(x_1=Yes\mid Healthy)P(x_2=No\mid Healthy)P(x_3=No\mid Healthy)P(x_4=Yes\mid Healthy)=$$

$$\frac{2}{5}\cdot\frac{3}{5}\cdot\frac{3}{5}\cdot\frac{1}{5}=\frac{18}{625}$$

$$P(Ill\mid\mathbf{x})=\frac{P(Ill)P(\mathbf{x}\mid Ill)}{P(Ill)P(\mathbf{x}\mid Ill)+P(Healthy)P(\mathbf{x}\mid Healthy)}=$$

$$\frac{\frac{1}{2}\cdot\frac{36}{625}}{\frac{1}{2}\cdot\frac{36}{625}+\frac{1}{2}\cdot\frac{18}{625}}=\frac{36}{36+18}=\frac{2}{3}\approx 0.6667$$

$$P(Healthy\mid\mathbf{x})=1-P(Ill\mid\mathbf{x})=\frac{1}{3}\approx 0.3333$$

Consequently, this person is predicted to be ill.

c) A person with a running nose and reddened skin: **x**=(Yes,No,Yes,No).

$$P(\mathbf{x}\mid Ill)=P(x_1=Yes\mid Ill)P(x_2=No\mid Ill)P(x_3=Yes\mid Ill)P(x_4=No\mid Ill)=$$

$$\frac{3}{5}\cdot\frac{3}{5}\cdot\frac{3}{5}\cdot\frac{3}{5}=\frac{81}{625}$$

$$P(\mathbf{x}\mid Healthy)=$$

$$P(x_1=Yes\mid Healthy)P(x_2=No\mid Healthy)P(x_3=Yes\mid Healthy)P(x_4=No\mid Healthy)=$$

$$\frac{2}{5}\cdot\frac{3}{5}\cdot\frac{2}{5}\cdot\frac{4}{5}=\frac{48}{625}$$

$$P(Ill\mid\mathbf{x})=\frac{P(Ill)P(\mathbf{x}\mid Ill)}{P(Ill)P(\mathbf{x}\mid Ill)+P(Healthy)P(\mathbf{x}\mid Healthy)}=$$

$$\frac{\frac{1}{2}\cdot\frac{81}{625}}{\frac{1}{2}\cdot\frac{81}{625}+\frac{1}{2}\cdot\frac{48}{625}}=\frac{81}{81+48}=\frac{27}{43}\approx 0.6279$$

$$P(Healthy\mid \mathbf{x})=1-P(Ill\mid \mathbf{x})=\frac{16}{43}\approx 0.3721$$

Consequently, this person is predicted to be ill.

Exercise 2:

In this exercise C=4, V={*Heal*, *Infl*, *Salm*, *Bow*}, and D=4. Consequently, for every $y\in V$ we have:

$$P(y\mid \mathbf{x})=\frac{P(y)P(\mathbf{x}\mid y)}{\sum_{v\in V}P(v)P(\mathbf{x}\mid v)}$$

$$P(\mathbf{x}\mid y)=\prod_{d=1}^{4}P(x_d\mid y)$$

From the class frequencies in the training set we obtain the class prior probabilities:

$$P(Heal)=\frac{1}{7}$$

$$P(Infl)=\frac{2}{7}=P(Salm)=P(Bow)$$

On the other hand:

$$P(x_1\mid y)=\frac{n'+\frac{4}{3}}{n+4}$$

$$P(x_2\mid y)=\frac{n'+2}{n+4}$$

$$P(x_3\mid y)=\frac{n'+2}{n+4}$$

$$P(x_4\mid y)=\frac{n'+2}{n+4}$$

Now we can compute the class probabilities for the test data:

a) A person with high fever: **x**=(high,No,No,No).

$$P(\mathbf{x}\mid Heal)=$$
$$P(x_1=high\mid Heal)P(x_2=No\mid Heal)P(x_3=No\mid Heal)P(x_4=No\mid Heal)=$$
$$\frac{4}{15}\cdot\frac{3}{5}\cdot\frac{3}{5}\cdot\frac{3}{5}=\frac{36}{625}$$

$$P(\mathbf{x}\mid Infl)=$$
$$P(x_1=high\mid Infl)P(x_2=No\mid Infl)P(x_3=No\mid Infl)P(x_4=No\mid Infl)=$$

$$\frac{7}{18}\cdot\frac{4}{6}\cdot\frac{4}{6}\cdot\frac{3}{6}=\frac{336}{3888}=\frac{7}{81}$$

$$P(\mathbf{x}\mid Salm)=$$

$$P(x_1=high\mid Salm)P(x_2=No\mid Salm)P(x_3=No\mid Salm)P(x_4=No\mid Salm)=$$

$$\frac{7}{18}\cdot\frac{3}{6}\cdot\frac{2}{6}\cdot\frac{4}{6}=\frac{168}{3888}=\frac{7}{162}$$

$$P(\mathbf{x}\mid Bow)=$$

$$P(x_1=high\mid Bow)P(x_2=No\mid Bow)P(x_3=No\mid Bow)P(x_4=No\mid Bow)=$$

$$\frac{4}{18}\cdot\frac{2}{6}\cdot\frac{2}{6}\cdot\frac{4}{6}=\frac{64}{3888}=\frac{2}{243}$$

$$P(Heal\mid\mathbf{x})=\frac{P(Heal)P(\mathbf{x}\mid Heal)}{\sum_{v\in V}P(v)P(\mathbf{x}\mid v)}=$$

$$\frac{\frac{1}{7}\cdot\frac{36}{625}}{\frac{1}{7}\cdot\frac{36}{625}+\frac{2}{7}\cdot\frac{7}{81}+\frac{2}{7}\cdot\frac{7}{162}+\frac{2}{7}\cdot\frac{2}{243}}=$$

$$\frac{\frac{36}{4375}}{\frac{1205}{25306}}=\frac{849}{4913}\approx 0.1728$$

$$P(Infl\mid\mathbf{x})=\frac{P(Infl)P(\mathbf{x}\mid Infl)}{\sum_{v\in V}P(v)P(\mathbf{x}\mid v)}=$$

$$\frac{\frac{2}{7}\cdot\frac{7}{81}}{\frac{1}{7}\cdot\frac{36}{625}+\frac{2}{7}\cdot\frac{7}{81}+\frac{2}{7}\cdot\frac{7}{162}+\frac{2}{7}\cdot\frac{2}{243}}=$$

$$\frac{\frac{2}{81}}{\frac{1205}{25306}}=\frac{937}{1807}\approx 0.5185$$

$$P(Salm\mid\mathbf{x})=\frac{P(Salm)P(\mathbf{x}\mid Salm)}{\sum_{v\in V}P(v)P(\mathbf{x}\mid v)}=$$

$$\frac{\frac{2}{7}\cdot\frac{7}{162}}{\frac{1}{7}\cdot\frac{36}{625}+\frac{2}{7}\cdot\frac{7}{81}+\frac{2}{7}\cdot\frac{7}{162}+\frac{2}{7}\cdot\frac{2}{243}}=$$

$$\frac{\frac{1}{81}}{\frac{1205}{25306}}=\frac{937}{3614}\approx 0.2593$$

$$P(Bow \mid \mathbf{x}) = \frac{P(Bow)P(\mathbf{x} \mid Bow)}{\sum_{v \in V} P(v)P(\mathbf{x} \mid v)} =$$

$$\frac{\frac{2}{7} \cdot \frac{2}{243}}{\frac{1}{7} \cdot \frac{36}{625} + \frac{2}{7} \cdot \frac{7}{81} + \frac{2}{7} \cdot \frac{7}{162} + \frac{2}{7} \cdot \frac{2}{243}} =$$

$$\frac{\frac{4}{1701}}{\frac{1205}{25306}} = \frac{313}{6338} \approx 0.0494$$

Consequently, this person is predicted to suffer from influenza.

b) A person who suffers from vomiting and shivering: **x**=(No,Yes,No,Yes).

$$P(\mathbf{x} \mid Heal) =$$

$$P(x_1 = No \mid Heal)P(x_2 = Yes \mid Heal)P(x_3 = No \mid Heal)P(x_4 = Yes \mid Heal) =$$

$$\frac{7}{15} \cdot \frac{2}{5} \cdot \frac{3}{5} \cdot \frac{2}{5} = \frac{84}{625}$$

$$P(\mathbf{x} \mid Infl) =$$

$$P(x_1 = No \mid Infl)P(x_2 = Yes \mid Infl)P(x_3 = No \mid Infl)P(x_4 = Yes \mid Infl) =$$

$$\frac{4}{18} \cdot \frac{2}{6} \cdot \frac{4}{6} \cdot \frac{3}{6} = \frac{96}{3888} = \frac{2}{81}$$

$$P(\mathbf{x} \mid Salm) =$$

$$P(x_1 = No \mid Salm)P(x_2 = Yes \mid Salm)P(x_3 = No \mid Salm)P(x_4 = Yes \mid Salm) =$$

$$\frac{4}{18} \cdot \frac{3}{6} \cdot \frac{2}{6} \cdot \frac{2}{6} = \frac{48}{3888} = \frac{1}{81}$$

$$P(\mathbf{x} \mid Bow) =$$

$$P(x_1 = No \mid Bow)P(x_2 = Yes \mid Bow)P(x_3 = No \mid Bow)P(x_4 = Yes \mid Bow) =$$

$$\frac{4}{18} \cdot \frac{4}{6} \cdot \frac{2}{6} \cdot \frac{2}{6} = \frac{64}{3888} = \frac{2}{243}$$

$$P(Heal \mid \mathbf{x}) = \frac{P(Heal)P(\mathbf{x} \mid Heal)}{\sum_{v \in V} P(v)P(\mathbf{x} \mid v)} =$$

$$\frac{\frac{1}{7} \cdot \frac{84}{625}}{\frac{1}{7} \cdot \frac{84}{625} + \frac{2}{7} \cdot \frac{2}{81} + \frac{2}{7} \cdot \frac{1}{81} + \frac{2}{7} \cdot \frac{2}{243}} =$$

$$\frac{527}{882} \approx 0.5975$$

$$P(Infl\mid \mathbf{x})=\frac{P(Infl)P(\mathbf{x}\mid Infl)}{\sum_{v\in V}P(v)P(\mathbf{x}\mid v)}=$$

$$\frac{\frac{2}{7}\cdot\frac{2}{81}}{\frac{1}{7}\cdot\frac{84}{625}+\frac{2}{7}\cdot\frac{2}{81}+\frac{2}{7}\cdot\frac{1}{81}+\frac{2}{7}\cdot\frac{2}{243}}=$$

$$\frac{537}{2446}\approx 0.2195$$

$$P(Salm\mid \mathbf{x})=\frac{P(Salm)P(\mathbf{x}\mid Salm)}{\sum_{v\in V}P(v)P(\mathbf{x}\mid v)}=$$

$$\frac{\frac{2}{7}\cdot\frac{1}{81}}{\frac{1}{7}\cdot\frac{84}{625}+\frac{2}{7}\cdot\frac{2}{81}+\frac{2}{7}\cdot\frac{1}{81}+\frac{2}{7}\cdot\frac{2}{243}}=$$

$$\frac{446}{4063}\approx 0.1098$$

$$P(Bow\mid \mathbf{x})=\frac{P(Bow)P(\mathbf{x}\mid Bow)}{\sum_{v\in V}P(v)P(\mathbf{x}\mid v)}=$$

$$\frac{\frac{2}{7}\cdot\frac{2}{243}}{\frac{1}{7}\cdot\frac{84}{625}+\frac{2}{7}\cdot\frac{2}{81}+\frac{2}{7}\cdot\frac{1}{81}+\frac{2}{7}\cdot\frac{2}{243}}=$$

$$\frac{179}{2446}\approx 0.0732$$

Consequently, this person is predicted to be healthy.

Exercise 3:

In this exercise C=2, V={Buy, $NoBuy$}, and D=3. Consequently, for every $y\in V$ we have:

$$P(y\mid \mathbf{x})=\frac{P(y)P(\mathbf{x}\mid y)}{P(Buy)P(\mathbf{x}\mid Buy)+P(NoBuy)P(\mathbf{x}\mid NoBuy)}$$

$$P(\mathbf{x}\mid y)=\prod_{d=1}^{3}P(x_d\mid y)$$

From the class frequencies in the training set we obtain the class prior probabilities:

$$P(Buy)=\frac{9}{14}$$

$$P(NoBuy)=\frac{5}{14}$$

On the other hand:

$$P(x_1 \mid y) = \frac{n' + \frac{5}{3}}{n+5}$$

$$P(x_2 \mid y) = \frac{n' + \frac{5}{2}}{n+5}$$

$$P(x_3 \mid y) = \frac{n' + \frac{5}{3}}{n+5}$$

Now we can compute the class probabilities for the test sample, **x**=(Medium,Yes,Fair):

$$P(\mathbf{x} \mid Buy) =$$
$$P(x_1 = Medium \mid Buy)P(x_2 = Yes \mid Buy)P(x_3 = Fair \mid Buy) =$$
$$\frac{17}{42} \cdot \frac{17}{28} \cdot \frac{20}{42} = \frac{728}{6221}$$

$$P(\mathbf{x} \mid NoBuy) =$$
$$P(x_1 = Medium \mid NoBuy)P(x_2 = Yes \mid NoBuy)P(x_3 = Fair \mid NoBuy) =$$
$$\frac{11}{30} \cdot \frac{7}{20} \cdot \frac{8}{20} = \frac{77}{1500}$$

$$P(Buy \mid \mathbf{x}) = \frac{P(Buy)P(\mathbf{x} \mid Buy)}{P(Buy)P(\mathbf{x} \mid Buy) + P(NoBuy)P(\mathbf{x} \mid NoBuy)} =$$
$$\frac{\frac{9}{14} \cdot \frac{728}{6221}}{\frac{9}{14} \cdot \frac{728}{6221} + \frac{5}{14} \cdot \frac{77}{1500}} = \frac{3135}{3899} = 0.8041$$

$$P(NoBuy \mid \mathbf{x}) = \frac{P(NoBuy)P(\mathbf{x} \mid NoBuy)}{P(Buy)P(\mathbf{x} \mid Buy) + P(NoBuy)P(\mathbf{x} \mid NoBuy)} =$$
$$\frac{\frac{5}{14} \cdot \frac{77}{1500}}{\frac{9}{14} \cdot \frac{728}{6221} + \frac{5}{14} \cdot \frac{77}{1500}} = \frac{764}{3899} \approx 0.1959$$

Exercise 4:

In this exercise C=2, V={*Diabetes*, *NoDiabetes*}. Consequently, for every $y \in V$ we have:

$$P(y \mid \mathbf{x}) = \frac{P(y)P(\mathbf{x} \mid y)}{P(Diabetes)P(\mathbf{x} \mid Diabetes) + P(NoDiabetes)P(\mathbf{x} \mid NoDiabetes)}$$

The a priori class probabilities are:

$$P(Diabetes) = 10.6$$

$$P(NoDiabetes) = 89.4$$

a) The observed data is **x**=(Positive). Then we have:

$$P(\mathbf{x} \mid Diabetes) = 0.99$$

$$P(\mathbf{x} \mid NoDiabetes) = 0.06$$

Now we can compute the probability to suffer from diabetes:

$$P(Diabetes \mid \mathbf{x}) =$$

$$\frac{P(Diabetes)P(\mathbf{x} \mid Diabetes)}{P(Diabetes)P(\mathbf{x} \mid Diabetes) + P(NoDiabetes)P(\mathbf{x} \mid NoDiabetes)} =$$

$$\frac{10.6 \cdot 0.99}{10.6 \cdot 0.99 + 89.4 \cdot 0.06} \approx 0.6617$$

b) After the second test the observed data is **x**=(Positive,Negative). Then we have:

$$P(\mathbf{x} \mid Diabetes) = 0.99 \cdot 0.01 = 0.0099$$

$$P(\mathbf{x} \mid NoDiabetes) = 0.06 \cdot 0.94 = 0.0564$$

Now we can recompute the probability to suffer from diabetes:

$$P(Diabetes \mid \mathbf{x}) =$$

$$\frac{P(Diabetes)P(\mathbf{x} \mid Diabetes)}{P(Diabetes)P(\mathbf{x} \mid Diabetes) + P(NoDiabetes)P(\mathbf{x} \mid NoDiabetes)} =$$

$$\frac{10.6 \cdot 0.0099}{10.6 \cdot 0.0099 + 89.4 \cdot 0.0564} \approx 0.0204$$

Exercise 5:

a)

$$Accuracy = \frac{n_{11} + n_{22}}{n_{11} + n_{12} + n_{21} + n_{22}} =$$

$$\frac{182 + 245}{182 + 31 + 12 + 245} \approx 0.9085$$

b)

$$TP = n_{11} = 182$$

$$TN = n_{22} = 245$$

$$FP = n_{12} = 31$$

$$FN = n_{21} = 12$$

c)

$$Precision = \frac{182}{182 + 31} \approx 0.8545$$

$$Fallout = \frac{31}{31 + 245} \approx 0.1123$$

$$Recall = \frac{182}{182 + 12} \approx 0.9381$$

$$F - measure = 2\frac{0.8545 \cdot 0.9381}{0.8545 + 0.9381} = 0.8944$$